



Center for  
K–12 Assessment  
& Performance Management

*An independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.*

## Exploratory Seminar:

Measurement Challenges Within  
the Race to the Top Agenda

December 2009

## RESPONSES TO SESSION 1: MEASURING GROWTH

This policy brief is based on reactions by Wendy M. Yen (Educational Testing Service) and Mike Kane (Educational Testing Service) to the Session 1 presentations at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009. Download a copy of the final papers written by Dr. Yen and Dr. Kane, as well as other papers presented at the seminar, at <http://www.k12center.org/publications.html>.

### WENDY M. YEN, EDUCATIONAL TESTING SERVICE

As Betebenner and Linn pointed out, the results of the measurement of growth in high-stakes accountability systems will depend on the types of data used and the analyses of them. Growth models can be a valuable addition to status models when reviewing progress, but they should not replace status models entirely. Nor should one particular growth model be mandated. Whether using status or growth, Betebenner and Linn rightly indicated that the measurement properties of the procedure should be commensurate with the requirements of the analysis.

### Vertical Scales

Sometimes a vertical scale is used in measuring academic growth. The assumption behind vertical scales is that the content of adjacent grades is well articulated, but if a state's standards/curriculum have large subareas that are not designed to be taught hierarchically, a vertical scale cannot be expected to produce sensible results. Even if the Common Core of Standards produces vertical articulation, as grades get further apart and the tested content diverges, the entire concept of equivalence of scores and score units does not have a logical basis and is, in fact, untestable. The results of high-stakes summative assessments are scrutinized in minute detail and small differences can have big policy implications. Vertical scales, which are based on untestable assumptions, cannot be demonstrated to have the accuracy needed for wide-ranging applications in high-stakes accountability testing. Regression models can mitigate the problems with vertical scales because they can measure growth with minimal assumptions and compare the performance of students who start at the same place, that is, with the same scores at the end of the previous grade.

*Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.*

## **Uses of Accountability Testing**

Accountability testing can stimulate thinking on alignment of standards, curricula, and assessments, and it can help us define and focus on what we want students to learn. Accountability testing can help us see what change is taking place and what is not. For example, tracking of annual testing results in California has shown sustained improvement in the percentage of proficient students but virtually no change in the gap between White students and students of color. Historical accountability testing information can help us evaluate the difficulty of goals and the resources needed to meet them. It can help create stretch goals that are attainable, by identifying schools with the top improvement records and using them as examples. On the negative side, unintended negative consequences can occur with accountability systems, particularly if their performance goals are unlikely to be obtained.

## **Recommendations**

These thoughts lead to my recommendations for a revised accountability system under the Race to the Top initiative:

- Broaden the measures of achievement and success beyond one summative, high-stakes assessment; that is, examine multiple measures of performance.
- Develop goals for improvement that are challenging but attainable (historical information is important in setting such goals).
- Measure and equate accurately (changes in tests or testing conditions can have large consequences).
- Encourage growth measures that do not rely on untestable assumptions and support both absolute and normative interpretations.
- Be cautious in attributing causation to results, be they status or growth results.
- Emphasize accessible communication to users (Colorado’s visual reporting on its website is a positive example).
- Innovate responsibly (do not be overconfident about innovations nor oversell them).

The current development of next-generation K-12 assessments undoubtedly will produce new ideas and innovations. When psychometric innovations are implemented in high-stakes testing, it is incumbent upon the testing community to demonstrate before the implementation that the measurement properties of the system, particularly the equivalence and comparability of scores, are sufficient for their intended use.

## **MIKE KANE, EDUCATIONAL TESTING SERVICE**

The presentations by Pellegrino and Betebenner and Linn focused on asking the key question, Growth in what? Presumably, we want students who start as novices to develop in competence and/or expertise in significant domains, keeping in mind that it is good for society to have people with different sets of skills.

Learning progressions or learning trajectories offer a promising way to analyze growth because they can be specified at different levels of detail. They can tie assessment results to important outcomes, they are qualitative, they are built around the conceptions of growth, and they can remind us that growth in competence and expertise is not one-dimensional.

Answering the question as to how much growth we have is tricky because growth usually is not one-dimensional, and a simple how much suggests a one-dimensional quantitative scale. In most areas, the learning progressions tend to be curriculum and classroom-bound, and even within a particular classroom, students will vary in what they attend to and what they are good at. I am not confident that we have a complete enough set of variables to support value-added measures of how much growth.

Lastly, even though the shift from less punitive consequences to more descriptive information suggested by Betebenner and Linn would be challenging, it is a battle worth fighting.

### **Major Points of Open Discussion**

Participants who had listened to the presentations and responses seemed to agree that there would always be tensions over which model (status or growth) to use, but that the critical issue is how the information is used and by whom. Some argued that local people, particularly teachers, need greater capacity to analyze and use assessment data. A countervailing view was that it is passing the buck to ask teachers to decide if the data is worth using, especially because the presentations expressed major reservations about the growth models. There is an ethical dilemma in pushing assessment systems on local educators when there are still suppositions about the systems, said one participant, but the overall goal should be to encourage good decisionmaking on the part of teachers and principals so that they move in the right direction. Easier said than done, though, it was pointed out, because there is no model yet for helping educators use and integrate local data with other data. Policymakers especially deal with such a range of issues and organizations that they cannot possibly know all they should, so assessment experts and researchers have to give policymakers understandable measurement results, which often means results are pared down. Whether using descriptive or causal information, knowing how to describe the information is critically important. Even when using descriptive data, people must understand the relationship between instruction and growth.

There was general endorsement of the perspective offered on the science of learning and its ability to reach agreement on what is to be learned. While learning progressions seem more suitable for diagnosing, they can also inform the development of summative assessments so that they reflect continuity from one grade to another. Currently, however, common assessments are being imposed on differing curriculum and pacing. All the more reason, a presenter added, to link assessment to instruction, not in the lockstep model of the French education system but in a model that at least provides a better idea of where learning is headed.

Changing the paradigm for assessment systems seems overwhelming, but one participant suggested that changes made in another sector—welfare policy—offer a strategy. Small experiments in individual states build up over time to large-scale adoption of reforms.

### **For More Information**

For more information on this subject, please see the papers by Dr. Betebenner and Dr. Linn, Dr. Kane, Dr. Pellegrino, and Dr. Yen:

Betebenner, D. W., & Linn, R.L. (2010). *Achievement: Issues of measurement, longitudinal data analysis, and accountability*. Retrieved from <http://www.k12center.org/publications.html>.

Kane, M. (2010). *Comments on growth in achievement*. Retrieved from <http://www.k12center.org/publications.html>.

Pellegrino, J. W. (2010). *The design of an assessment system for the race to the top: A learning sciences perspective on issues of growth and measurement*. Retrieved from <http://www.k12center.org/publications.html>.

Yen, W. M. (2010). *Measuring student growth with large-scale assessments in an education accountability system*. Retrieved from <http://www.k12center.org/publications.html>.